Diabetes and machine learning

Created by:
Chris Duran
Kimberly Kruel
Mistie McClure
Geoff McDaniel

Our Goal and Methodology

Our objective is to determine the effectiveness of various modeling techniques in accurately predicting diabetes presence based on features derived from survey questions.

At a high level, this project has the potential to address the diabetes challenge by leveraging predictive modeling to potentially identify individuals at risk of developing diabetes based on their responses to survey questions and other relevant features.
By accurately predicting diabetes risk, the project can hopefully help those at risk seek proper testing or lifestyle changes.

Fast Facts on Diabetes

Diabetes

- **Total:** 38.4 million people have diabetes (11.6% of the US population)
- **Diagnosed:** 29.7 million people, including 29.4 million adults
- **Undiagnosed:** 8.7 million people (22.8% of adults are undiagnosed)

Prediabetes

- **Total:** 97.6 million people aged 18 years or older have prediabetes (38.0% of the adult US population)
- **65 years or older:** 27.2 million people aged 65 years or older (48.8%) have prediabetes

Our Data Sets

Behavioral Risk Factor Surveillance System Public health surveys of people from 2011-2015

https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system

Acknowledgements:
This dataset was released by the CDC (Centers for Disease Control and Prevention)

Diabetes Health Indicators Dataset https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Acknowledgements:
This dataset was released by the CDC (Centers for Disease Control and Prevention)

DATA PREPARATION

Overview of Data Collections

- CDC Datasets
- Survey results 2011-2015
- AWS Storage
- Sample size = 100,000 per year

Consolidation & Transformation

- Consolidate to shared features
- Replace bad data answers
- Create date references
- Transform select features to categories

Review

- Correlation
- Review remaining feature set
- Pivot

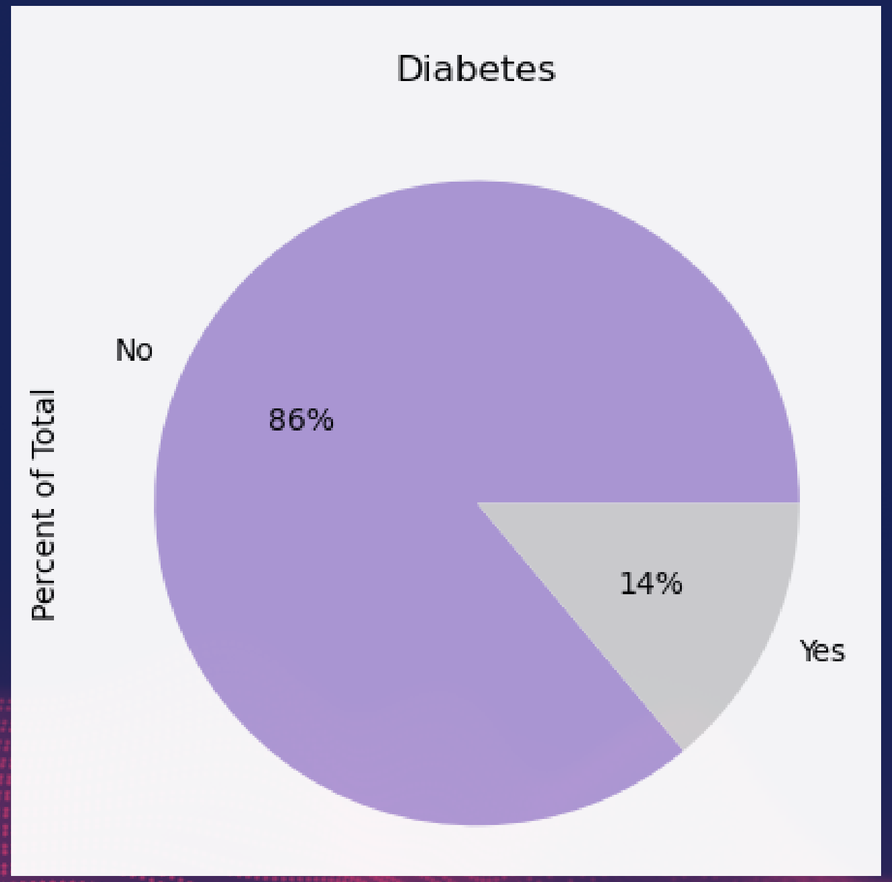
EXPLORATORY DATA ANALYSIS

- Dataset Assembly:
  - Assembled datasets from various years for comprehensive analysis.
  - Pivoted to approach of just 2015 data due to difference in questions and data provided.
  - Focused analysis on the year 2015, narrowing down to relevant datasets.
- Filtering and Standardization:
  - Conducted rigorous filtering to extract diabetes-related variables.
  - Standardized dataset to enhance manageability and consistency.
- Dataframe Characteristics:
  - The final dataset contains 21 key health indicators, focusing on aspects like general health, physical health, mental health, and diabetes status.

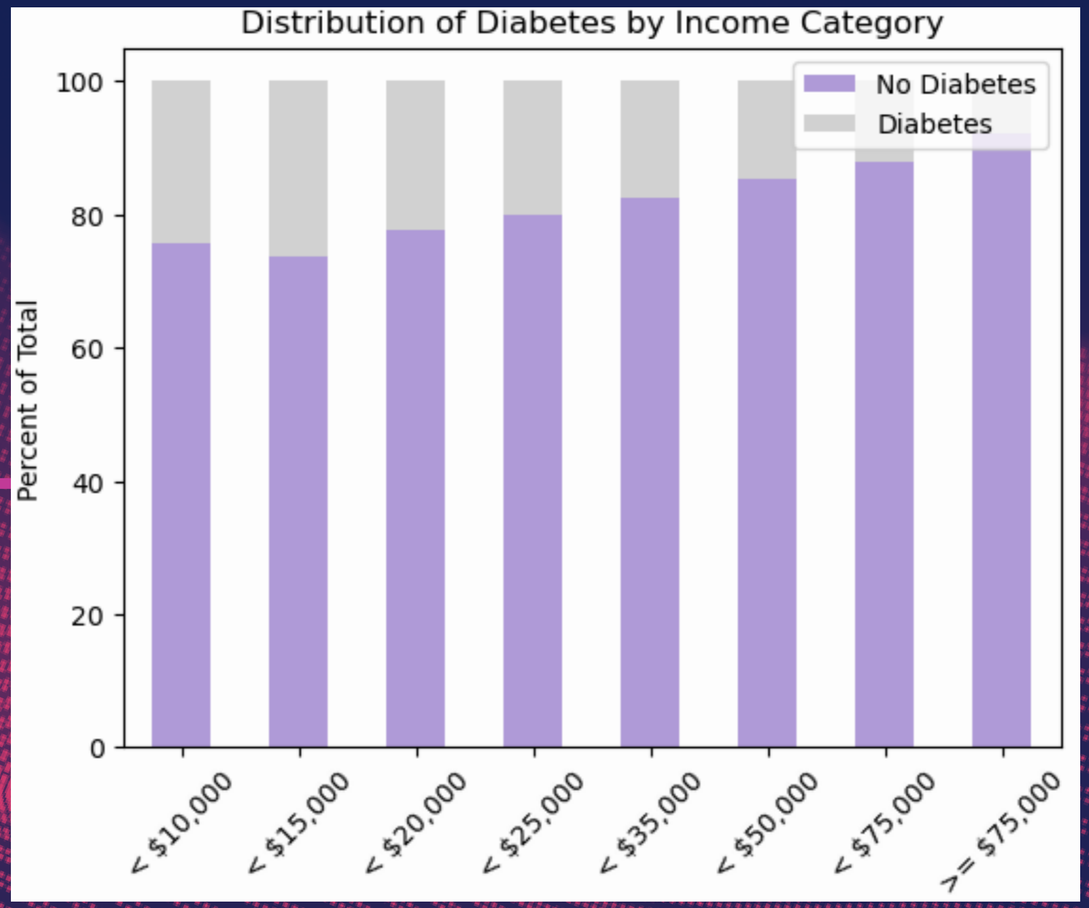
Top Complications Related to Diabetes

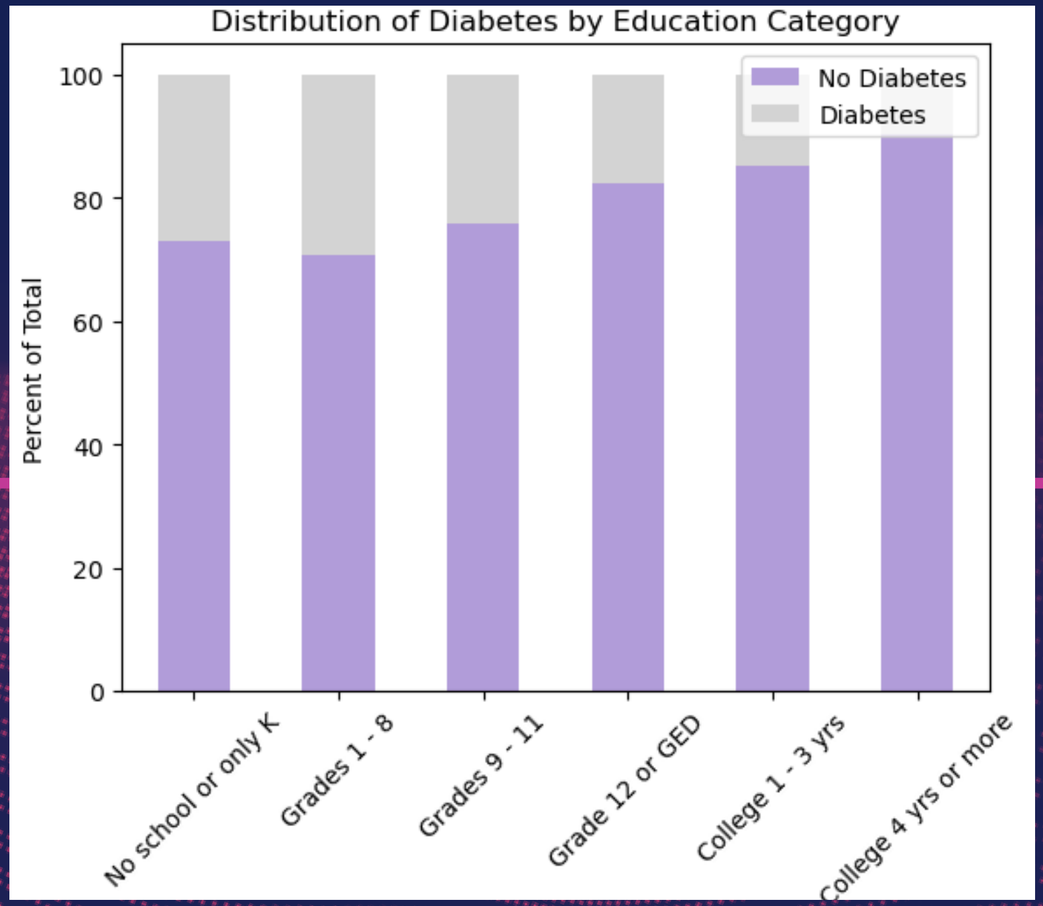
1. Heart disease: Diabetes doubles an individual's risk of heart disease which is the leading cause of death in the US.
2. Cancer: Diabetes is related to a higher risk of certain types of cancer due to the effect high blood sugar has on certain organs.
3. Viral: Heart and clotting complications from any virus are higher in individuals with poorly managed diabetes.
4. Stroke: Due to the inherent nature of diabetes weakening effect on the blood vessels, individuals living with diabetes are 1.5-2 times more likely to suffer from a stroke and are more likely to have poor post stroke outcomes.

The CDC estimates that 20% of individuals living with diabetes are unaware of their condition.



Education, Income, and Diabetes





The data establishes a clear gradient in the relationship diabetes has with income and education.

Visualizing the data side by side shows that lower levels of education and income are almost identical in their correlation to a higher rate of diabetes.

It would be worth reviewing for causation to determine if enhanced education for dietary and lifestyle could be implemented at an earlier age in curriculum or as part of public assistance programs.

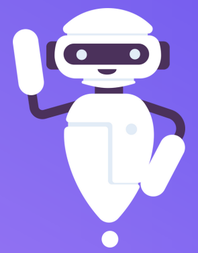
The top prevention and treatments for diabetes are dietary adjustments and physical activity.

OUR MODELING

MODELS WE USED

- Logistic Regression
- Random Forest
- Decision Tree
- SVC (Support Vector Classifier)
- XGBoost
- K-Nearest Neighbors (KNN)
- AdaBoost

HYPERTUNED MODELS



MODEL 01

Logistic Regression

Train Accuracy:
0.8630085146641439
Test Accuracy:
0.8650898770104068

MODEL 02

ADAboost

Train Accuracy:
0.8602806685588142
Test Accuracy:
0.8618259224219489

MODEL 03

XGBoost

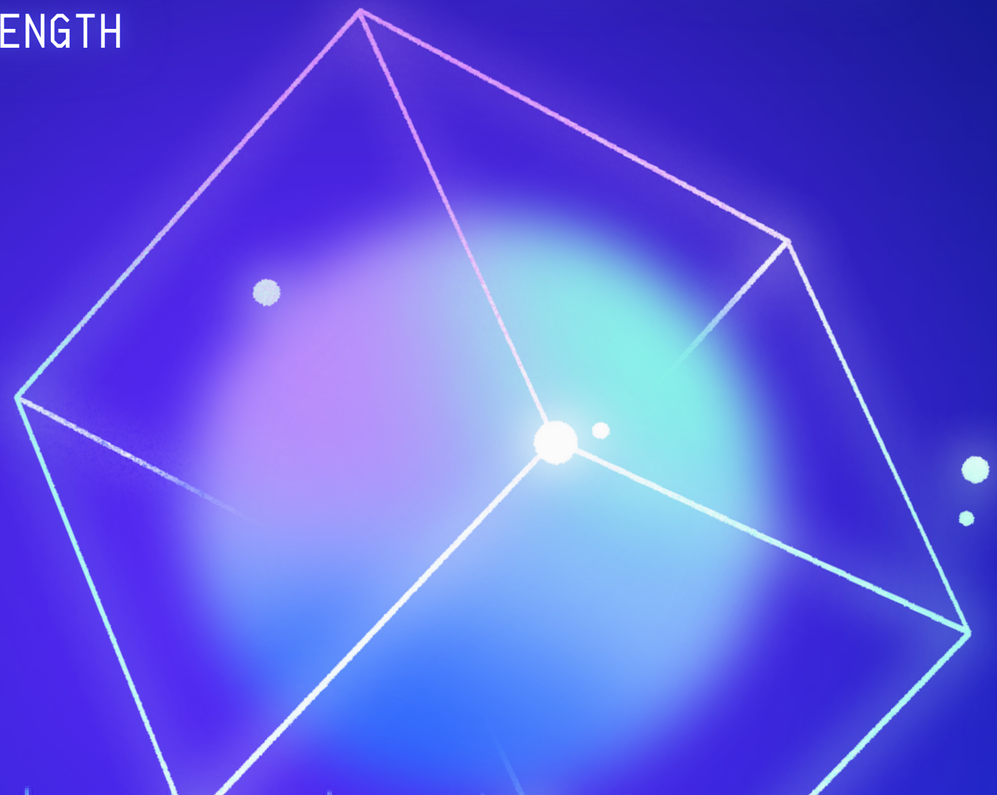
Train Accuracy:
0.8673394302533375
Test Accuracy:
0.8672500788394828

LOGISITC REGRESSION HYPERTUNING

- MAX ITERATIONS (MAX_ITER): EXPLORED VALUES - 100, 200, 300, 400, 500
- SOLVER: CHOICES - LBFGS, LIBLINEAR, SAG, SAGA
- INVERSE REGULARIZATION STRENGTH (C): RANGE - 1 TO 499

ADABOOST HYPERTUNING

- NUMBER OF ESTIMATORS (N_ESTIMATORS): EXPLORED VALUES - 50, 100, 200, 300, 400, 500
- LEARNING RATE (LEARNING_RATE): RANGE - 1 TO 499



XGBOOST HYPERTUNING

01

Number of Estimators (n_estimators): Explored values - 50, 100, 200, 300, 400, 500

02

Max Depth (max_depth): Range - 1 to 4

03

Max Leaves (max_leaves): Range - 1 to 4

04

Min Child Weight (min_child_weight): Range - 1 to 4

05

Subsample Ratio (subsample): Range - 0.1 to 0.9 (step of 0.1)

KEY FINDINGS

What do the models tell us about the data?

- Uniform Accuracy of 86% across the models
- Well-structured data (expected from Kaggle)
- Features have consistent impact
- High false positive predictions

Which model(s) performed the best?

- Logistic Regression
- Tree Boosting (Ada & XG)

What are some important uses of this info?

- Deployment as an app to both Healthcare providers and individuals which may be at risk
- Deployment as an LLM Agent

	precision	recall	f1-score	support
negative	0.88	0.98	0.93	54657
positive	0.57	0.16	0.25	8763
accuracy			0.87	63420
macro avg	0.72	0.57	0.59	63420
weighted avg	0.84	0.87	0.83	63420

THANK YOU


WHERE ARE
OUR GRADES